ARTICLES FROM THE SCIP CONFERENCE

Using SAS PROC NL MIXED to fit item response theory models

CHING-FAN SHEU

DePaul University, Chicago, Illinois

and

CHENG-TE CHEN, YA-HUI SU, and WEN-CHUNG WANG

National Chung Cheng University, Min Hsiung, Taiwan

Researchers routinely construct tests or questionnaires containing a set of items that measure personality traits, cognitive abilities, political attitudes, and so forth. Typically, responses to these items are scored in discrete categories, such as points on a Likert scale or a choice out of several mutually exclusive alternatives. Item response theory (IRT) explains observed responses to items on a test (questionnaire) by a person's unobserved trait, ability, or attitude. Although applications of IRT modeling have increased considerably because of its utility in developing and assessing measuring instruments, IRT modeling has not been fully integrated into the curriculum of colleges and universities, mainly because existing general purpose statistical packages do not provide built-in routines with which to perform IRT modeling. Recent advances in statistical theory and the incorporation of those advances into general purpose statistical software such as the Statistical Analysis System (SAS) allow researchers to analyze measurement data by using a class of models known as generalized linear mixed effects models (McCulloch & Searle, 2001), which include IRT models as special cases. The purpose of this article is to demonstrate the generality and flexibility of using SAS to estimate IRT model parameters. With real data examples, we illustrate the implementations of a variety of IRT models for dichotomous, polytomous, and nominal responses. Since SAS is widely available in educational institutions, it is hoped that this article will contribute to the spread of IRT modeling in quantitative courses.

In assessment and evaluation research, a person's ability or attitude is usually measured by a test or a questionnaire. Answers to the questions can be dichotomous (*yes* or *no*) or polytomous (points on a Likert scale). Item response theory (IRT) models the relationship, in probabilistic terms, between a person's response to an item and his or her standing on the construct being measured by the scale. In education, medicine, and social research, IRT models play an increasingly important role in the construction and validation of measurement scales.

Currently, a behavioral scientist wishing to perform IRT analysis can choose one (or several) specialized software packages, such as BIGSTEPS/WINSTEPS (Linacre

& Wright, 1999), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1997), and ConQuest (Wu, Adams, & Wilson, 1998). Frequently, more than one computer program has to be used because these programs, free or commercially available, do not implement the same set of IRT models, nor do they employ the same parameter estimation method, not to mention sharing a common user interface. None of the specialized software, moreover, is particularly easy to learn or use (Hays, Morales, & Reise, 2000). Clearly, it is more efficient for researchers and students alike to conduct IRT modeling on a general purpose computing platform, since most of them are familiar with and have ready access to such a platform.

From the statistical point of view, IRT is a collection of nonlinear mixed effects models that link observed responses to latent variables. They are mixed effects models because the parameters representing the characteristics (e.g., difficulty) of the items are fixed, whereas the person parameter is considered random. They are nonlinear because the link functions between the response variable and the parameters are allowed to be nonlinear. Agresti, Booth, Hobert, and Caffo (2000), in a survey of

This research was supported in part by National Science Council of Taiwan Grant NSC 92-2811-H-194-001 and by a Paid Leave Program granted to the first author from the University Research Council of DePaul University. The hospitality and support of the Department of Psychology at National Chung Cheng University is gratefully acknowledged. We thank David Allbritton and two anonymous reviewers for a number of helpful comments and suggestions. Correspondence concerning this article should be addressed to C.-F. Sheu, Department of Psychology, DePaul University, 2219 North Kenmore Ave., Chicago, IL 60614-3522 (e-mail: csheu@condor.depaul.edu).

random effects modeling of binary and count data, included the Rasch model as an early example of mixed logistic models. Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) further explicated the connections between nonlinear mixed models and common IRT models. A consequence of such connections between the two classes of models is that multilevel software packages such as HLM (Raudenbush, Bryk, Cheong, & Congdon, 2000) and MLwiN (Rasbash et al., 2000), although not designed to fit IRT models to data, can nevertheless be adapted to estimate IRT model parameters through ingenious reformulation (Kamata, 2001). However, all of these computer programs share the same shortcoming of being special purpose software.

The purpose of this article is to illustrate the generality and flexibility of using the Statistical Analysis System (SAS) to estimate the parameters of IRT models. The popularity of SAS as a general purpose statistical software package makes this approach to IRT model estimation attractive to a wide audience in academic and industrial settings. The generality of this approach is demonstrated by implementing a variety of IRT models for dichotomous, polytomous, and nominal responses in a single routine, PROC NL MIXED. The ease with which a new model can be implemented in SAS by simply altering a few lines of previously written code for a different IRT model testifies to the flexibility of this approach. A recent study by Smits, De Boeck, and Verhelst (2003) in which the estimations of two different compartmental IRT models were compared concluded that the SAS approach is more flexible than the use of the special purpose software MIRID CML.

This article is organized as follows. The next section will introduce and describe the Rasch model (Rasch, 1960) and the three-parameter logistic models (Birnbaum, 1968) for binary data. The former will then be illustrated with responses to items on Euclidean geometry (Woodhouse, 1991); the latter will be illustrated with a data set from the *Profile of American Youth* (U.S. Department of Defense, 1982). The marginal maximum likelihood estimation will be briefly explained. The third section will deal with the graded response (Samejima, 1969), the rating scale (Andrich, 1978), and the partial credit (Masters, 1982) models for polytomous responses. These models will be illustrated with responses to statements regarding capital punishment (Roberts, 1995). The generalized partial credit model (Muraki, 1992) will be used to fit responses to the items of a science test with mixed numbers of answer categories (Adams, Doig, & Rosier, 1991). In the fourth section, the nominal categories model (Bock, 1972) will be illustrated with an analysis of life satisfaction data (Davis, 1975). The fifth section will give a brief explanation of goodness-of-fit statistics commonly used in IRT practice to diagnose items or persons that deviate from the model's expectation. Throughout the article, we will discuss each of the IRT models by first presenting the model for the data at hand, commenting on the corresponding SAS statements

implementing the model, and then comparing the results of parameter estimates with published reports, if available, or with those obtained from specialized IRT software packages. In the last section, we will draw some general conclusions on using SAS to perform IRT modeling.

DICHOTOMOUS RESPONSES

The simplest case of scoring a test occurs when responses to items are judged as either correct or incorrect—in the case of items on a questionnaire, as *yes* or *no*. Consider the responses of a sample of 150 individuals to nine multiple-choice items on two-dimensional Euclidean geometry (Woodhouse, 1991). The responses were coded 1 for correct answers and 0 for incorrect. This data set is available at the Centre for Multilevel Modelling on the Web at <http://multilevel.ioe.ac.uk/intro/datasets.html> under the link "Item response data" in "Data files." The name of the zip file for download is *starter.zip*, and we shall refer to this data set as *starter*.

Four observations may be made regarding test data such as these. (1) Given a group of individuals with varying levels of ability in two-dimensional Euclidean geometry, we would expect that the higher a person's ability, the higher the likelihood that this person will answer a large number of items on the test correctly. (2) The number of persons who could be expected to answer a difficult item correctly should be smaller than those who would answer an easier item on the test correctly. (3) The nine multiple-choice items, if well constructed, should test nothing other than an individual's ability in two-dimensional Euclidean geometry, and each item should provide independent information on the individual's ability. (4) The responses of an individual are not influenced by the responses of the others.

The Rasch model, or the one-parameter logistic model (Rasch, 1960), took the observations above as assumptions. A formal discussion of these assumptions can be found in Sijtsma and Molenaar (2002).

The Rasch Model

For the dichotomously scored answers of 150 persons to nine items that appeared to measure the same ability in two-dimensional Euclidean geometry in the *starter* data, the Rasch model specifies the probability of a person $i = 1, 2, \dots, 150$ answering correctly item $j = 1, 2, \dots, 9$ as

$$P(X_{ij} = 1 | \theta_i, b_j) = \frac{1}{1 + \exp[-(\theta_i - b_j)]},$$

where X_{ij} is a Bernoulli random variable representing person i 's response to item j , b_j is the difficulty of item j , and θ_i represents person i 's ability on a single continuum. The probability of a respondent's answering an item correctly is, therefore, a function of the difference between the respondent's ability in two-dimensional Euclidean geometry and the difficulty of the item. The person pa-

rameters are assumed to be independent and identically distributed normal distributions with a mean of zero and a variance of σ^2 . In other words, the person parameters are random effects, as opposed to the item parameters, which are fixed effects.

Parameter estimation. A variety of procedures have been used to estimate item parameters of IRT models. These procedures can be categorized as based either on the method of maximum likelihood or on some other method. In the likelihood-based category are three commonly used procedures: joint maximum likelihood, conditional maximum likelihood, and marginal maximum likelihood methods. The first estimates the person and item parameters jointly in an iterative fashion, the second eliminates the person parameters by conditioning on a sufficient statistic for θ , and the third maximizes the marginal likelihood by first integrating out the person parameters and using the first- and second-order derivatives to obtain the item parameter estimates. The estimates of person parameters can then be obtained from the now known values of item parameters. The last method is implemented by SAS in the NLMIXED procedure (SAS Institute, 2000; Wolfinger, 1999), and we will briefly explain it below. Readers who are interested in the details of estimation methods can consult van der Linden and Hambleton (1997). The first step in using the method of maximum likelihood to estimate parameters is to specify the likelihood function, which is the marginal density function of the observed data viewed as a function of the model parameters. Under the local independence assumption, which all IRT models share and which means that a person's response to an item is not influenced by his or her responses to other items on the same test, the likelihood, for instance, of the starter data under the Rasch model is as shown at the bottom of the page, where x_{ij} is the response of person i for item j and $\phi(\theta)$ is the density function of a normal distribution with a variance parameter σ^2 .

The marginal maximum likelihood estimation requires maximizing the likelihood or, equivalently, the logarithm of the likelihood, which, in turn, requires integrating the joint probability function of the responses with respect to the person distributions (random effects). For practical purposes, numerical, instead of the often intractable analytical, integrations are carried out using Gaussian quadrature (Abramowitz & Stegun, 1972) formulas to obtain density weights at a number of evaluation points. For nonlinear mixed models, the NLMIXED procedure directly fits the specified model by maximizing an approximation to the likelihood integrated over the random effects. An adaptive version of Gauss–Hermite quadrature

is used to approximate the likelihood, and the default maximization routine is a dual quasi-Newton algorithm (Pinheiro & Bates, 1995).

In practice, responses to test items by a sample of subjects are laid out in a person (row) $\times$ item (column) data frame. For example, the file starter0.dat is a 150 (person) $\times$ 9 (item) matrix whose elements are the responses of the subjects. To use the NLMIXED procedure, the data must be organized in a format in which each row is a person's response to a single item, with its corresponding indicator (dummy) coding for the item. Listing 1 presents the responses of the first and last subjects to nine test items in the starter example in what is called a *long* format. The first dozen lines preceding PROC NLMIXED in Listing 2 recode the usual IRT data to the long format, using a call to PROC TRANSPOSE followed by a DATA step.

To convert a different data file, the user simply edits the relevant variable names to reflect the changes of file name or number of items. The ensuing program listings assume that appropriate changes to command lines preceding PROC NLMIXED have been made to ensure that the data have been converted to the proper format. The last three lines in Listing 2 before the final RUN statement save predicted response probabilities and item and person parameter estimates to SAS internal data locations indicated by the corresponding variable names. Their contents will later be recalled to calculate fit statistics, to be discussed in the Goodness of Fit section. The rest of the SAS commands in Listing 2 illustrate the NLMIXED statements for estimating the person and item difficulty parameters of the Rasch model for the starter data.

The program was executed on a Pentium M Processor, 1300-MHz notebook computer with 512 MB of RAM running on Windows XP. The same computer was used to perform all the subsequent analyses in this article. The version of SAS was 8.02. The log files reported running times under 35 min for the examples presented here. Most runs were completed within a few minutes.

To implement an IRT model in NLMIXED, the main step is to explicitly state the likelihood of the model in a clear and structured way. This is achieved by defining only one factor of the likelihood (the response), using indicator variables for the items, separating the different response categories in the polytomous case, and considering the subjects as units of random observations. The following discussion will go into the details concerning the NLMIXED statements for the Rasch model.

The PROC NLMIXED statement invokes the procedure and inputs the starter data set. The QPOINTS option specifies the number of quadrature points to use in approximating the likelihood with Gauss–Hermite quadrature.

$$\prod_{i=1}^{150} \left(\int_{\theta} \prod_{j=1}^9 \left\{ \frac{1}{1 + \exp[-(\theta_i - b_j)]} \right\}^{x_{ij}} \left\{ \frac{\exp[-(\theta_i - b_j)]}{1 + \exp[-(\theta_i - b_j)]} \right\}^{(1-x_{ij})} \phi(\theta) d\theta \right)$$

It is recommended that the number of quadrature points be increased until the results of the parameter estimates are stabilized. Our experience showed that for IRT models, it often requires using about 20 to 30 quadrature points. In most cases, we set QPOINTS to 25. Setting a high number of quadrature points, at the cost of longer computing time, does not appear to decrease the standard errors of parameter estimates. However, the effect of the number of quadrature points on the accuracy of analysis is a complicated matter, since it depends as much on the structures of the models as on the configurations of data. A detailed simulation study examining this thorny issue can be found in Lesaffre and Spiessens (2001).

The PARMS statement defines parameters and sets initial values for them. The initial values of the parameter must be supplied by the user. An accurate guess on what the starting values might be can speed up convergence.

Within the NLMIXED procedure, we define the conditional probabilities for dichotomous responses as a function of the linear predictor *eta*. The response probabilities are then specified according to the cumulative logistic function. The MODEL statement defines the response variable and its conditional distribution, given the random effect, which is a respondent's ability parameter. The RANDOM statement specifies the random effect, *u*, as a normal distribution with a mean of μ and a variance of σ^2 . Agresti et al. (2000) recommended specifying the variance in terms of the standard deviation *sigma* in NLMIXED, because it can improve the stability of numerical solutions in cases in which the estimated variance is very close to zero. The SUBJECT argument is set to the variable PERSON, indicating that the random effect changes according to the value of the subject identification number.

Table 1 compares the parameter estimates obtained from NLMIXED and those from the WINSTEPS computer program (Linacre & Wright, 1999), with standard errors of the estimates. The WINSTEPS program automatically adjusts the mean of item estimates to zero. Thus, we set the mean of the random effect μ in Listing 2 to the value 1.155 estimated by WINSTEPS in order to facilitate comparison. The two sets of estimates are virtu-

ally the same, although WINSTEPS uses the joint maximum likelihood method and PROC NLMIXED uses the marginal likelihood method to estimate parameters. As compared with the values reported by NLMIXED, a trade-off is found in WINSTEPS between the smaller standard errors for the item parameter estimates and the larger standard error for the estimate of σ .

The Three-Parameter Logistic Model

As an extension of the Rasch model for dichotomous responses, Birnbaum's (1968) three-parameter logistic model expresses the probability of a correct response to item *j* from subject *i* as

$$P(X_{ij} = 1 | \theta_i, a_j, b_j, c_j) = c_j + \frac{(1 - c_j)}{1 + \exp[-1.7a_j(\theta_i - b_j)]},$$

where X_{ij} is the response (1 if correct and 0 otherwise), θ_i is the latent ability of the subject *i*, and a_j , b_j , and c_j are parameters characterizing item *j*: a_j reflects the rate of change in the proportion of correct responses to the item as a function of an individual's latent ability, b_j reflects the level of θ at which a subject has a 50% chance of correctly answering the item (without guessing and assuming that all items have the same discrimination power), and c_j reflects the minimum probability of a correct response from guessing alone. The constant 1.7 is a scaling factor to bring the cumulative logistic function numerically in line with the cumulative standard normal distribution function. The person parameters θ_i are random effects and take the form of independent and identically distributed normal distributions with a mean of zero and a variance of σ^2 . Notice that the two-parameter logistic model is obtained by setting the guessing parameters c_j in the three-parameter logistic model to zero. (*Two-parameter* or *three-parameter* here refers to the number of distinct sets of item parameters, not the actual number of parameters in the model.)

We will illustrate the fit of the three-parameter logistic model to data from the *Profile of American Youth* (U.S. Department of Defense, 1982), a survey of the aptitudes of a national probability sample of Americans 16–23 years of age in July 1980. The data reported in Mislevy (1985) contains responses of 776 subjects to four items from the arithmetic reasoning test of the Armed Services Vocational Aptitude Battery (ASVAB), Form 8A.

In addition to the (full) three-parameter logistic model in which the guessing parameters are assumed to be different for different items, we also will consider a restricted version of the model in which a common minimum probability of a correct response from guessing alone is assumed for all four items. Table 2 displays the parameter estimates obtained from fitting the restricted version of the model to the army data, using both the PROC NLMIXED and the BILOG-MG programs (Zimowski et al., 1996). The SAS implementation of the (full) three-parameter logistic model to estimate the person and item parameters is shown in Listing 3.

Table 1
Parameter Estimates From PROC NLMIXED and WINSTEPS
for the Starter Data Set, With Standard Errors

Parameter	NLMIXED		WINSTEPS	
	Est.	SE	Est.	SE
b1	-1.75	0.34	-1.71	0.28
b2	-0.55	0.24	-0.58	0.21
b3	-0.19	0.23	-0.22	0.19
b4	-0.11	0.23	-0.14	0.19
b5	0.13	0.23	0.10	0.19
b6	0.21	0.23	0.18	0.18
b7	-0.71	0.25	-0.73	0.21
b8	0.95	0.21	0.95	0.18
b9	2.02	0.22	2.15	0.12
sigma	1.11	0.12	1.21	0.16

Table 2
Parameter Estimates From PROC NLMIXED and
BILOG-MG for Fitting the Three-Parameter Logistic Model
(With a Common Guessing Parameter for Items) to the
Army Data Set, With Standard Errors

Parameter	NLMIXED		BILOG-MG	
	Est.	SE	Est.	SE
a1	1.56	0.43	1.30	0.30
a2	0.88	0.18	0.88	0.15
a3	1.19	0.34	1.13	0.25
a4	3.36	1.18	1.63	0.55
b1	-0.02	0.09	-0.08	0.10
b2	0.27	0.11	0.21	0.12
b3	0.63	0.10	0.57	0.11
b4	0.75	0.10	0.77	0.11
c	0.20	0.03	0.18	0.04

The BOUND statement ensures that the values of slope parameters a_j are positive and that the values of guessing parameters c_j are between 0 and 1. For identifiability, the variance of the person parameter θ_i is set to 1. BILOG-MG also assumes that the random effects (person parameters) follow the standard normal distribution. In fitting complex models, it is often difficult to guess simultaneously the initial values of the many parameters involved. We resolved this by carrying out the estimation of the three-parameter logistic model in NLMIXED in two stages. In the first stage, we set all the slope parameters a_j equal to 1 and all the guessing parameters c_j to 0 and estimated only the location parameters b_j . (This is equivalent to fitting a one-parameter logistic model to the data.) These estimates were then used as starting values in the second stage of estimation. Table 2 shows that the parameter estimates obtained from PROC NLMIXED were quite similar to those from BILOG-MG, with the largest discrepancy occurring in the estimates for slope parameter a_4 . The difference was, however, not statistically significant. The standard error estimate of a_4 reported by PROC NLMIXED is also larger than that reported by BILOG-MG. It is noted that the estimates obtained from the BILOG-MG were not identical to those reported by Mislevy (1985) with BILOG (Mislevy & Bock, 1982).

NLMIXED reports the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion (Kass & Raftery, 1995), as well as the usual log-likelihood statistic (Read & Cressie, 1988), which can be used to compare the fit of such models as the one-parameter, two-parameter, and three-parameter logistic models for the same data. Table 3 shows the values of these model selection criteria for fitting the (hierarchical) logistic models to the army data. It appears that the results favor the three-parameter logistic model with a common guessing parameter as fitting the army data best. This finding is consistent with a single value of .2 for the guessing parameters reported in Mislevy (1985) and with our experience that if restraints are not imposed, the estimated values of the parameters in a complicated model are likely to increase without bound (Lord, 1975).

POLYTOMOUS RESPONSES

A polytomous item allows the respondent to endorse one of several possible response categories. For example, Roberts (1995) presented 245 subjects with a set of 24 statements on capital punishment, such as "Capital punishment may be wrong but it is the best preventative to crime," and asked them to express agreement or disagreement on a 6-point scale, where 1 = *strongly disagree*, 2 = *disagree*, 3 = *slightly disagree*, 4 = *slightly agree*, 5 = *agree*, and 6 = *strongly agree*. The data set is available at <http://www.education.umd.edu/EDMS/tutorials/data.html>. In some tests or questionnaires, the number of response categories may vary from item to item. Except for making the notation of IRT models slightly more complicated, allowing the number of response categories to vary in a test contributes little to the present discussion. Therefore, we will focus on tests or questionnaires in which all the items have the same number of response categories.

A variety of IRT models have been proposed for items with polytomous responses. Here, we shall demonstrate how to fit the capital punishment data with the graded response model (Samejima, 1969), the rating scale model (Andrich, 1978), and the partial credit model (Masters, 1982). The latter two models are extensions of the Rasch model for dichotomous responses, and their parameters can be estimated using WINSTEPS, which considers only models within the Rasch family. The graded response model, on the other hand, is not a member of the Rasch family. The MULTILOG program (Thissen, 1991) is used to fit the graded response model to the capital punishment data. For simplicity, we chose only responses to Statements 2, 13, 15, 16, and 19 of the original 24 statements in the following analysis. Because the response probabilities of all categories for an item must sum to 1, with six categories there are five *free* response probabilities. Thus, $k - 1$ steps (or thresholds) are introduced on the latent scale to model k category probabilities in polytomous IRT models.

Table 3
Values of Three Model Selection Criteria (-2 Log-Likelihood,
Akaike's Information Criterion [AIC], and the Bayesian
Information Criterion [BIC]) for Fitting Each of Four Models
(the One-Parameter [1-PL], the Two-Parameter [2-PL], and
2 Three-Parameter [3-PL] Logistic Models) to the
Army Data Set, Using PROC NLMIXED

Model	-2 Log-Likelihood	AIC	BIC	No. Parameters
1-PL	4032.3	4040.3	4059.0	4
2-PL	4005.5	4021.5	4058.7	8
3-PL	3985.4	4003.5	4045.3	9
3-PL	3979.8	4003.8	4059.6	12

Note—The guessing parameters in the first version of the 3-PL model are assumed to be the same for all four items, giving 9 parameters in total; in the second version, these parameters are assumed to be different for different items, yielding 12 parameters in total. The smaller the values of the AIC, BIC, and -2 log-likelihood, the better the model fit.

The Graded Response Model

The graded response model (Samejima, 1969) for polytomous items can be seen as an instance of random-effects cumulative logit (probit) models for ordinal data (Agresti et al., 2000). Sheu (2002) discussed using PROC NLMIXED to fit these models to repeated ordinal responses.

Let X_{ij} denote the response of a person i to a statement j on the 6-point rating scale of the capital punishment study. The quantity $P(X_{ij} \geq k)$ is the probability that the response takes on a value greater than or equal to a particular value $k = 1, \dots, 6$. By definition, the (cumulative) probability of responding in or above the lowest category is 1. The probability of a particular ordered categorical response is defined through the cumulative probabilities—for example,

$$P(X_{ij} = 3) = P(X_{ij} \geq 3) - P(X_{ij} \geq 4).$$

Thus, in terms of categorical response probabilities for item j by person i , the graded response model for the capital punishment data is given as follows:

$$P(X_{ij} = 1 | \theta_i, a_j, b_{j1}) = 1 - \frac{1}{1 + \exp[-a_j(\theta_i - b_{j1})]},$$

$$P(X_{ij} = k | \theta_i, a_j, b_{jk}) = \frac{1}{1 + \exp[-a_j(\theta_i - b_{j(k-1)})]} - \frac{1}{1 + \exp[-a_j(\theta_i - b_{jk})]},$$

for $k = 2, 3, 4$, and 5 ,

$$P(X_{ij} = 6 | \theta_i, a_j, b_{j5}) = \frac{1}{1 + \exp[-a_j(\theta_i - b_{j5})]},$$

where a_j is the slope parameter for item j and $b_{jk} = b_{j1} + d_{j(k-1)}$, $d_{j0} = 0$, $d_{j1}, \dots, d_{j4} \geq 0$, $\sum_{k=1}^5 d_{jk} = 0$, $j = 1, \dots, 5$. The d_{jk} are the step size parameters for item j . These parameters are item dependent—that is, different for each item on the questionnaire. The SAS implementation of the graded response model for the responses to 5 items of capital punishment data is presented in Listing 4. The BOUNDS statement ensures that the values of slope parameters a_j and step size parameters d_{jk} are all positive. The linear components inside the exponential function, *eta1* through *eta5*, are specified according to the model. The categorical response probabilities are then specified according to the cumulative logit model. The multinomial distribution is not directly supported by the MODEL statement in the NLMIXED procedure. Instead, a general likelihood function is specified using the GENERAL statement. The likelihood based on the probability (p) is checked to see whether it is numerically too close to zero, then converted to the log likelihood (ll). The log likelihood is set to a large negative value ($-1E100$) if the likelihood is close to zero. As usual, person pa-

Table 4
Parameter Estimates of the Graded Response Model
From PROC NLMIXED and MULTILOG for the
Five Statements in the Capital Punishment Data Set,
With Standard Errors

Parameter	NLMIXED		MULTILOG	
	Est.	SE	Est.	SE
a1	1.94	0.25	1.95	0.23
a2	2.41	0.31	2.42	0.26
a3	1.57	0.21	1.58	0.21
a4	2.60	0.34	2.61	0.28
a5	1.88	0.24	1.88	0.23
b11	-0.22	0.11	-0.24	0.11
b12	0.55	0.11	0.53	0.12
b13	1.13	0.14	1.11	0.16
b14	1.69	0.18	1.67	0.22
b15	2.17	0.24	2.14	0.28
b21	-0.38	0.10	-0.40	0.11
b22	0.47	0.10	0.45	0.10
b23	1.18	0.13	1.16	0.14
b24	1.49	0.15	1.47	0.17
b25	1.89	0.19	1.86	0.22
b31	-0.49	0.13	-0.51	0.14
b32	0.23	0.12	0.21	0.14
b33	0.91	0.14	0.88	0.17
b34	1.44	0.18	1.42	0.21
b35	2.16	0.26	2.13	0.28
b41	-0.49	0.10	-0.50	0.10
b42	0.24	0.09	0.22	0.09
b43	0.84	0.11	0.82	0.10
b44	1.44	0.14	1.41	0.15
b45	2.06	0.20	2.03	0.23
b51	-0.49	0.12	-0.51	0.13
b52	0.43	0.11	0.41	0.12
b53	0.97	0.13	0.94	0.15
b54	1.62	0.18	1.59	0.20
b55	2.25	0.25	2.23	0.29

rameters are defined as independent and identically distributed normal random variables with a constant standard deviation σ . The value of σ was set to 1 in order to facilitate results comparisons with the MULTILOG program, Version 7.02, which assumes that the person parameters follow the standard normal distribution.

Table 4 compares the parameter estimates of the graded response model obtained from the NLMIXED procedure with those from the MULTILOG program. The estimate values were virtually identical. The largest discrepancy was 0.03, which was very small.

The Rating Scale Model

In a rating scale model for polytomous responses, the response categories are scored so that the total score of a respondent represents a rating of the person's location on a latent scale. This model assumes that the scoring of the response categories within an item is constant across all items in the questionnaire.

Let X_{ij} be the response of individual i on item j to the five selected statements in the capital punishment data.

Andrich's (1978) rating scale model specifies the probability of response k , $k = 1, \dots, 6$ as

$$P(X_{ij} = k | \theta_i, b_j, d_l) = \frac{\exp \left[\sum_{l=0}^{k-1} (\theta_i - b_j - d_l) \right]}{\sum_{m=1}^6 \exp \left[\sum_{l=0}^{m-1} (\theta_i - b_j - d_l) \right]},$$

where d_l , $l = 0, \dots, k-1$, is the location of l th step (cut point) relative to the item's location and $\sum_{l=0}^5 d_l = 0$ with $d_0 = 0$. The quantity $\sum_{l=0}^0 (\theta_i - b_j - d_l)$ is also, by definition, zero. The numerator of a particular categorical response probability is defined as a linear combination of person, item, and step parameters inside the exponential function. The response probability of a category is then specified as the ratio of the corresponding numerator over the sum of all numerator terms. Since the denominator includes the step parameters for all the categorical responses, the model captures the process of endorsing the most suitable category by simultaneously considering the locations of all the thresholds (cut points) separating the response categories. The SAS implementation of the rating scale model is shown in Listing 5. The ESTIMATE statement is used to obtain an estimate of d_5 , the fifth step parameter, whose value is constrained by the fact that the sum of all the step parameters must equal zero. This form of constraint was chosen in order to facilitate comparison of parameter estimates, shown in Table 5, with the program WINSTEPS.

The Partial Credit Model

The partial credit model (Masters, 1982) extends the Rasch model for binary responses to pairs of adjacent categories in a sequence of ordered responses. It can also been seen as a generalization of the rating scale model, in that the step sizes for items on a test are allowed to vary. Thus, for the five-item capital punishment data on a 6-point scale, there are 4 step parameters to estimate

under the rating scale model, whereas, under the partial credit model, 20 step parameters are to be estimated. The latter model expresses the probability of person i endorsing k response category on item j as

$$P(X_{ij} = k | \theta_i, b_j, d_{jl}) = \frac{\exp \left[\sum_{l=0}^{k-1} (\theta_i - b_j - d_{jl}) \right]}{\sum_{m=1}^6 \exp \left[\sum_{l=0}^{m-1} (\theta_i - b_j - d_{jl}) \right]},$$

where d_{jl} , $l = 0, \dots, k-1$, is the location of l th step relative to the location of item j and for each item j , $\sum_{l=0}^5 d_{jl} = 0$ with $d_{j0} = 0$ following identifiability constraints. This means that the attractiveness of a response category for an item is measured only relative to the reference category of that item. The quantity $\sum_{l=0}^0 (\theta_i - b_j - d_{jl})$ is also, by definition, zero for each item j . In the partial credit model, the probability of each category of response for an item depends on the difficulty of all the thresholds for the item. The SAS implementation of the partial credit model is shown in Listing 6. It will not escape the reader's observation that Listings 4, 5, and 6 bear a strong similarity to one another.

The partial credit model implemented in the NLMIXED procedure produced parameter estimates that are very similar to those obtained from WINSTEPS, as is shown in Table 6.

The Generalized Partial Credit Model

For all the statements in the capital punishment example, the respondents were asked to express agreement or disagreement on a 6-point scale. In many applications of IRT, however, it is common to find items on a test containing different numbers of answer categories. For example, chap. 4 of the ConQuest manual presents an eight-item test in which three items are scored onto three performance levels and the remaining five items are scored onto four levels.

Items on tests with mixed numbers of response categories may not possess the same discriminating power. This may be construed as motivation to generalize the Rasch polytomous item response models, to account for the degree to which categorical responses vary among items as the ability level changes in the population. A case in point is the generalized partial credit model developed by Muraki (1992).

We will illustrate this model with an example containing the responses of 515 students to a test of science concepts related to the Earth and space, reported in Adams et al., (1991). The same data set with eight test items is available from the sample data files of the ConQuest software (Wu et al., 1998). In these data, the capital letters A, B, C, D, E, F, W, and X represent the different kinds of responses that students gave to these test items. For each item, these letters are scored to indicate the level of quality of the response. For example, responses to Items 2, 3, 4, and 6 are scored as follows:

Table 5
Parameter Estimates of the Rating Scale Model
From PROC NLMIXED and WINSTEPS for the
Five Statements in the Capital Punishment
Data Set, With Standard Errors

Parameter	NLMIXED		WINSTEPS	
	Est.	SE	Est.	SE
b1	0.94	0.10	0.94	0.06
b2	0.88	0.10	0.86	0.06
b3	0.66	0.09	0.61	0.06
b4	0.75	0.09	0.71	0.06
b5	0.77	0.10	0.71	0.06
d1	-0.67	0.10	-1.12	0.08
d2	-0.34	0.10	-0.34	0.08
d3	0.15	0.12	0.23	0.09
d4	0.40	0.15	0.51	0.11
d5	0.45	0.16	0.72	0.15
sigma	0.90	0.07	1.05	0.20

Table 6
Parameter Estimates of the Partial Credit Model
From PROC NLMIXED and WINSTEPS for the
Five Statements in the Capital Punishment
Data Set, With Standard Errors

Parameter	NLMIXED		WINSTEPS	
	Est.	SE	Est.	SE
b1	0.94	0.10	0.89	0.06
b2	0.88	0.10	0.79	0.06
b3	0.66	0.09	0.62	0.06
b4	0.75	0.09	0.78	0.06
b5	0.77	0.10	0.76	0.06
d11	-0.52	0.18	-0.91	0.17
d12	-0.25	0.23	-0.25	0.18
d13	0.07	0.28	0.13	0.21
d14	0.60	0.36	0.68	0.26
d15	0.10	0.37	0.34	0.34
d21	-0.78	0.18	-1.21	0.18
d22	-0.33	0.20	-0.31	0.17
d23	0.90	0.31	0.98	0.22
d24	0.09	0.39	0.19	0.26
d25	0.11	0.35	0.35	0.32
d31	-0.40	0.19	-0.85	0.18
d32	-0.50	0.22	-0.50	0.18
d33	0.20	0.16	0.28	0.19
d34	0.17	0.29	0.28	0.22
d35	0.54	0.29	0.78	0.29
d41	-0.79	0.19	-1.29	0.18
d42	-0.54	0.22	-0.55	0.17
d43	0.02	0.24	0.06	0.19
d44	0.54	0.30	0.66	0.24
d45	0.80	0.35	1.11	0.37
d51	-0.86	0.18	-1.33	0.18
d52	-0.08	0.22	-0.08	0.18
d53	-0.23	0.26	-0.15	0.19
d54	0.53	0.30	0.65	0.24
d55	0.64	0.34	0.52	0.34
sigma	0.92	0.08	1.05	0.20

Item 2: (A, B, C, W, X) → (3, 2, 1, 0, 0)

Item 3: (A, B, C, D, E, F, W, X) → (3, 2, 2, 1, 1, 0, 0, 0)

Item 4: (A, B, C, W, X) → (2, 1, 0, 0, 0)

Item 6: (A, B, W, X) → (2, 1, 0, 0).

For simplicity, we will use only the four items above to illustrate the implementation of the NLMIXED procedure for estimating parameters of the generalized partial credit model. For tests in which item responses have mixed numbers of categories, grouping items with the same number of categories together in either an ascending or descending order makes programming statements easier to follow and to debug, if necessary. We, therefore, will switch the order of items in the data, so that Items 4, 6, 2, and 3 in the original data become Items 1, 2, 3, and 4 in the new data.

The generalized partial credit model expresses the probability of person i endorsing k response category on item j as

$$P(X_{ij} = k | \theta_i, a_j, b_j, d_{jl}) = \frac{\exp \left[\sum_{l=0}^k a_j (\theta_i - b_j - d_{jl}) \right]}{\sum_{m=0}^3 \exp \left[\sum_{l=0}^m a_j (\theta_i - b_j - d_{jl}) \right]},$$

where d_{jl} , $l = 0, \dots, k$ is the location of l th step relative to the location of item j and, for each item j , $\sum_{l=0}^3 d_{jl} = 0$ with $d_{j0} = 0$ following identifiability constraints. The quantity $\sum_{l=0}^0 a_j (\theta_i - b_j - d_{jl})$ is also, by definition, zero for each item j .

Except for the changes in counting indices, to reflect the changes from the 6-point scale (ranges from 1 to 6) of the capital punishment data to the 0, 1, 2, and 3 scoring categories in the present example, the sole difference in the equation above from that given by the partial credit model is the introduction of the slope parameter a_j for each item j . In other words, one could view the generalized partial credit model as an extension of the two-parameter logistic model (Birnbaum, 1968) for dichotomous responses to polytomous responses. The slope parameter is usually assumed to range from zero to infinity. We note that since the step parameter d_{jl} is defined for each item j , the mixed number of answering categories for items on a test does not, by itself, pose a problem for the partial credit model.

Parameters for the generalized partial credit model can be estimated using the program PARSCALE (Muraki & Bock, 1997) or the NLMIXED procedure in SAS. Listing 7 shows PROC NLMIXED statements for fitting the model to the responses of 515 students to four items in the science test. An inspection of Table 7 reveals that the estimated values of model parameters from SAS and PARSCALE are practically the same.

NOMINAL RESPONSES

With nominally scored items, respondents are confronted with items that admit responses in a number of mutually exclusive categories. No intrinsic ordering of these response categories is assumed.

Davis (1975) reported responses of 1,472 participants in the 1975 General Social Survey to three questions regarding degree of satisfaction with family, hobbies, and residence. Clogg (1979) trichotomized the the seven-category

Table 7
Parameter Estimates of the Generalized Partial Credit Model
From PROC NLMIXED and PARSCALE for the Four
Items in the Science Data Set, With Standard Errors

Parameter	NLMIXED		PARSCALE	
	Est.	SE	Est.	SE
a1	1.10	0.18	1.08	0.11
a2	1.05	0.19	1.11	0.15
a3	0.36	0.08	0.42	0.05
a4	1.64	0.34	1.47	0.12
b1	0.54	0.10	0.58	0.09
b2	0.48	0.10	0.46	0.09
b3	2.60	0.60	2.25	0.16
b4	0.94	0.10	1.00	0.08
d11	1.24	0.14	1.26	0.10
d21	0.88	0.11	0.83	0.10
d31	3.15	0.69	2.73	0.26
d32	1.42	0.53	1.21	0.28
d41	1.88	0.17	1.97	0.08
d42	-0.88	0.19	-0.93	0.12

responses in the original data and summarized the counts of respondents in a $3 \times 3 \times 3$ cross-classification. Thissen and Steinberg (1988) ignored the *more or less* nature of satisfaction and treated the responses as nominal in order to illustrate their analysis of life satisfaction data with the nominal categories model proposed by Bock (1972). A small segment of the data file, *lifesat.dat*, is presented in Listing 8 in the long format required by PROC NLMIXED.

The Nominal Categories Model

The nominal categories model (Bock, 1972) is based on an extension of the bivariate logistic distributions derived by Gumbel (1961). For three 3-category items, the probability of a respondent i providing a response $k = 1, 2, 3$ to item $j = 1$ (family), 2 (hobbies), 3 (residence) takes the following form:

$$P(X_{ij} = k | \theta_i) = \frac{\exp(a_{jk}\theta_i + c_{jk})}{\sum_{l=1}^3 \exp(a_{jl}\theta_i + c_{jl})},$$

in which θ_i represents the unobserved standard normal distribution of satisfaction for respondent i , and a_{jk} and c_{jk} are the location and scale parameters on which Bock (1972) imposed the arbitrary linear restriction $\sum_{k=1}^3 a_{jk} = \sum_{k=1}^3 c_{jk} = 0$ to anchor the latent scale. Thissen and Steinberg (1988) imposed further constraints on parameters $a_{11} = a_{12}$ and $a_{21} = a_{31}$ to obtain a version of the nominal model whose parameter estimates fitting the life satisfaction data were reported in their Table 8. After reparameterization using the ESTIMATE statement, the implementation of the same nominal model presented in Listing 9 fit the same data with almost exactly the same parameter estimates, shown in our Table 8.

GOODNESS OF FIT

Although the main scope of this article deals with the parameter estimation of IRT models, using NLMIXED, in practice it is also often useful to have diagnostic checks on how well the model fits the data. Tests of item fit can identify items that do not measure the same latent trait as the other items on a test or items that are ill defined. Tests of person fit can identify persons whose responses

deviate from the general pattern of responses in the sample. In this section, we will briefly explain the popular residual fit statistics (Wright & Masters, 1982) of the Rasch model, implemented in the software WINSTEPS (Linacre & Wright, 1999). Discussions of other diagnostic indices can be found in Drasgow, Levine, and Williams (1985). The starter data example is used to illustrate the computations of these fit statistics as implemented by PROC IML (Interactive Matrix Language) using output from PROC NLMIXED.

Recall that, for the Rasch model in the starter example, the observed response X_{ij} is a Bernoulli random variable representing person i 's answer to item j , where the number of respondents n_i is 150 and the number of items n_j is 9. The expected value of X_{ij} is p_{ij} , the probability of a respondent correctly answering an item. The variance of X_{ij} is $W_{ij} = p_{ij}(1 - p_{ij})$. The standardized residuals are defined as

$$Z_{ij} = \frac{X_{ij} - p_{ij}}{\sqrt{W_{ij}}}.$$

After fitting the model to a collection of responses, the standardized residuals can be estimated by substituting the predicted probability $\hat{p}_{ij}$ for p_{ij} in the formula for Z_{ij} . For each item j , an unweighted (outfit) mean square is the average of squared standardized residuals across persons,

$$U_j = \sum_{i=1}^{n_j} \frac{Z_{ij}^2}{n_j},$$

and a weighted (infit) mean square is the variance-weighted average of squared standardized residuals,

$$V_j = \frac{\sum_{i=1}^{n_j} W_{ij} Z_{ij}^2}{\sum_{i=1}^{n_j} W_{ij}}.$$

Analogously, an unweighted mean square and a weighted mean square can be calculated for each person i —respectively,

$$U_i = \sum_{j=1}^{n_j} \frac{Z_{ij}^2}{n_j},$$

and

$$V_i = \frac{\sum_{j=1}^{n_j} W_{ij} Z_{ij}^2}{\sum_{j=1}^{n_j} W_{ij}}.$$

These mean-square fit statistics have an expected value of one. It has been suggested that the values of infit or outfit statistics should lie between 0.8 and 1.2 when the model's expectation is consistent with the data (Wright, Linacre, Gustafson, & Martin-Löf, 1994). When the value of the statistic is smaller than 0.8, the item (or person) is not informative. When it is greater than 1.2, the item (or person) is noisy or does not conform to the latent trait. In practice,

Table 8
Parameter Estimates of the Nominal Categories Model
From PROC NLMIXED for the Life
Satisfaction Data, With Standard Errors

Item	Parameter	Est.	SE	Parameter	Est.	SE
Family	a1	-0.54	0.08	c1	-1.38	0.09
	a2	-0.54	0.08	c2	-0.34	0.07
	a3	1.08	0.16	c3	1.71	0.10
Hobbies	a1	-0.64	0.08	c1	-0.85	0.07
	a2	-0.25	0.06	c2	0.07	0.05
	a3	0.88	0.07	c3	0.78	0.05
Residence	a1	-0.64	0.08	c1	-0.84	0.07
	a2	-0.25	0.06	c2	0.26	0.05
	a3	0.89	0.07	c3	0.58	0.05

the unweighted mean square can be overly sensitive to extreme unexpected responses with item difficulty estimates away from the respondent's ability estimate, whereas the weighted mean square can be overly sensitive to nonextreme unexpected answers with item difficulty estimates near the respondent's ability estimate. Because the mean-square statistics vary from item to item and from sample to sample, they do not provide a significant test of whether the model fits the set of responses. The infit or outfit statistic, however, can be transformed into a standardized fit statistic (ZSTD) by a cube-root transformation:

$$t(y) = \left(\frac{1}{y^3} - 1 \right) \left(\frac{3}{\sqrt{\text{Var}(y)}} \right) + \left(\frac{\sqrt{\text{Var}(y)}}{3} \right),$$

where y is either the infit or the outfit statistic and $\text{Var}(y)$ is its variance. Using a critical interval of ± 2.58 to screen a person or item as misfitting the model has approximately 1% Type I error rate, on the basis of the standard normal distribution.

Listing 10 shows the implementation in PROC IML of the computations of the fit statistics above for examining the Rasch model fit to the starter data. The statements in Listing 10 are constructed to follow those shown in Listing 1. They can be readily adapted to implement fit statistics for polytomous item response models.

Table 9 shows that, for the starter data example, Item 8 is the only poor-fitting item, judging by the critical value of the ZSTD. On the other hand, the size of misfit is negligible, since the value of MNSQ for this item lies within the critical interval (0.8, 1.2). The fit statistics for all the persons in this example show satisfactory data-model fit. Their values are not reported.

CONCLUSIONS

Using data examples taken from published sources, we have demonstrated the flexibility of implementing IRT parameter estimation in SAS PROC NL MIXED. Estimation of new or more complex IRT models can be accomplished through the general programming statements in SAS. In typical applications of IRT models, more than parameter estimation is involved, albeit this is the most

important step in the process. To aid users in interpreting the features of a fitted model, some of the specialized IRT software routinely generate plots, such as person-item maps and item response characteristic curves (trace lines), and a myriad of diagnostic indices, such as infit and outfit statistics. To emulate, within a general purpose software package, all the functionalities provided by these computer programs will require programming expertise and effort beyond the scope of this article. Computing time may be another important limiting factor in the full-scale application of IRT models with SAS. We have observed that parameter estimation can take several hours to complete for data sets consisting of thousands of respondents and tests or questionnaires having more than 50 items, which are not unusual in practice. Nevertheless, we believe that widespread adoption of IRT in behavioral research will be greatly facilitated by the development of a user-friendly IRT module within a general purpose statistical software package such as SAS.

By subsuming the IRT models under the generalized linear mixed-effects models, we have also demonstrated the generality of using PROC NL MIXED in SAS to perform data analysis with these models. This generality applies equally to other general purpose software, such as S-Plus or R, the open-source implementation of the S language. These programs can also be used to fit the generalized mixed-effects models and, by extension, the IRT models discussed here. Instructors wishing to introduce IRT models in their courses can do so following the standard discussion on generalized linear models and using the same general purpose software to carry out parameter estimations. Formulating the recently developed multidimensional IRT models, such as the testlet model (Wainer, Bradlow, & Du, 2000) and the bundle model (Wilson & Adams, 1995), within the framework of the generalized mixed-effects models allows researchers to readily fit these models to data without having to resort to specialized software packages (Wang, Chen, & Sheu, 2004).

REFERENCES

- ABRAMOWITZ, M., & STEGUN, I. A. (EDS.) (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (9th printing). New York: Dover.
- ADAMS, R. J., DOIG, B. A., & ROSIER, M. (1991). *Science learning in Victorian schools: 1990* (ACER Research Monograph No. 41). Hawthorn, Victoria: Australian Council for Educational Research.
- AGRESTI, A., BOOTH, J. G., HOBERT, J. P., & CAFFO, B. (2000). Random effects modeling of categorical response data. *Sociological Methodology*, **30**, 27-80.
- AKAIKE, H. (1973). Information theory and an extension of maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akadémiai Kiadó.
- ANDRICH, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, **43**, 561-573.
- BIRNBAUM, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Table 9
Item Fit Statistics From Fitting the Rasch Model
to the Starter Data Example

Item	Measure	SE	Item Fit Statistics			
			Outfit MNSQ	Outfit ZSTD	Infit MNSQ	Infit ZSTD
1	-1.75	0.34	1.00	0.13	1.07	0.35
2	-0.55	0.25	0.67	-1.89	0.87	-1.05
3	-0.19	0.23	0.76	-1.68	0.88	-1.23
4	-0.11	0.23	0.71	-2.19	0.84	-1.67
5	0.13	0.22	0.74	-2.33	0.82	-2.20
6	0.21	0.22	0.81	-1.70	0.91	-1.10
7	-0.71	0.25	0.73	-1.35	0.92	-0.53
8	0.95	0.21	0.81	-2.75	0.84	-2.68
9	2.02	0.21	0.96	-0.29	0.89	-1.44

- BOCK, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29-51.
- CLOGG, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research*, **8**, 287-301.
- DAVIS, J. A. (1975). *Codebook for the spring 1975 general social survey*. Chicago: National Opinion Research Center.
- DRASGOW, F., LEVINE, M. V., & WILLIAMS, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical & Statistical Psychology*, **38**, 67-86.
- GUMBEL, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association*, **56**, 335-349.
- HAYS, R. D., MORALES, L. S., & REISE, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, **38**(Suppl. 9), I128-I142.
- KAMATA, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, **38**, 79-93.
- KASS, R. E., & RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.
- LESAFFRE, E., & SPIESSENS, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society: Series C*, **50**, 325-335.
- LINACRE, J. M., & WRIGHT, B. D. (1999). *A user's guide to Bigsteps/Winsteps*. Chicago: Mesa.
- LORD, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin 77-33). Princeton, NJ: Educational Testing Service.
- MASTERS, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.
- MCCULLOCH, C. E., & SEARLE, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.
- MISLEVY, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, **80**, 993-997.
- MISLEVY, R. J., & BOCK, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific.
- MURAKI, E. J. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159-176.
- MURAKI, E. J., & BOCK, R. D. (1997). *PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales* [Computer program]. Chicago: Scientific Software International.
- PINHEIRO, J. C., & BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational & Graphical Statistics*, **4**, 12-35.
- RASBACH, J., BROWNE, W., GOLDSTEIN, H., YANG, M., PLEWIS, I., HEALY, M., WOODHOUSE, G., DRAPER, D., LANGFORD, I., & LEWIS, T. (2000). *A user's guide to MLwiN, Version 2.1*. London: University of London, Institute of Education.
- RASCH, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Pædagogiske Institut.
- RAUDENBUSH, S. W., BRYK, A. S., CHEONG, Y. F., & CONGDON, R. (2000). *HLM5: Hierarchical linear and nonlinear modeling* [Computer program]. Chicago: Scientific Software International.
- READ, T. R. C., & CRESSIE, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.
- RIJMEN, F., TUERLINCKX, F., DE BOECK, P., & KUPPENS, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, **8**, 185-205.
- ROBERTS, J. S. (1995). Item response theory approaches to attitude measurement (Doctoral dissertation, University of South Carolina, Columbia). *Dissertation Abstracts International*, **56**, 7089B.
- SAMEJIMA, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- SAS INSTITUTE (2000). *SAS/STAT user's guide* (Version 8). Cary, NC: Author.
- SHEU, C.-F. (2002). Fitting mixed-effects models for repeated ordinal outcomes with the NLMIXED procedure. *Behavior Research Methods, Instruments, & Computers*, **34**, 151-157.
- SIJTSMA, K., & MOLENAAR, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- SMITS, D. J. M., DE BOECK, P., & VERHELST, N. D. (2003). Estimation of the MIRID: A program and an SAS-based approach. *Behavior Research Methods, Instruments, & Computers*, **35**, 537-549.
- THISEN, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory* [Computer software]. Chicago: Scientific Software International.
- THISEN, D., & STEINBERG, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, **3**, 385-395.
- U.S. DEPARTMENT OF DEFENSE (1982). *Profile of American youth*. Washington, DC: Author, Office of the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics.
- VAN DER LINDEN, W. J., & HAMBLETON, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- WAINER, H., BRADLOW, E. T., & DU, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). London: Kluwer.
- WANG, W.-C., CHEN, C.-T., & SHEU, C.-F. (2004, August). *Formulating multidimensional item response models using the SAS NLMIXED procedure*. Paper presented at the 28th International Congress of Psychology, Beijing.
- WILSON, M., & ADAMS, R. J. (1995). Rasch models for item bundles. *Psychometrika*, **60**, 181-198.
- WOLFINGER, R. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. *SUGI 24 Conference Proceedings*, Paper 287, Cary, NC: SAS Institute.
- WOODHOUSE, G. (1991). Multilevel item response models. In R. Prosser, J. Rasbash, & H. Goldstein (Eds.), *Data analysis with ML3* (pp. 79-88). London: University of London, Institute of Education.
- WRIGHT, B. D., LINACRE, J. M., GUSTAFSON, J. E., & MARTIN-LÖF, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, **8**, 370.
- WRIGHT, B. D., & MASTERS, G. N. (1982). *Rating scale analysis*. Chicago: Mesa.
- WU, M. L., ADAMS, R. J., & WILSON, M. R. (1998). *Generalized item response modelling software* [Computer software]. Melbourne: ACER Press.
- ZIMOWSKI, M. F., MURAKI, E., MISLEVY, R. J., & BOCK, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software.

LISTING 1
Responses of the First and Last Respondents to Nine Items of Euclidean Geometry,
Shown in a Long Format (i.e., One Response per Row)

1	1	0	0	0	0	0	0	0	0	1
1	0	1	0	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	0	1
1	0	0	0	1	0	0	0	0	0	1
1	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	0	1	0	0	0	1
1	0	0	0	0	0	0	1	0	0	1
1	0	0	0	0	0	0	0	1	0	1
1	0	0	0	0	0	0	0	0	1	1
1	0	0	0	0	0	0	0	0	1	0
150	1	0	0	0	0	0	0	0	0	1
150	0	1	0	0	0	0	0	0	0	1
150	0	0	1	0	0	0	0	0	0	1
150	0	0	0	1	0	0	0	0	0	1
150	0	0	0	0	1	0	0	0	0	1
150	0	0	0	0	0	1	0	0	0	1
150	0	0	0	0	0	0	1	0	0	1
150	0	0	0	0	0	0	0	1	0	1
150	0	0	0	0	0	0	0	0	1	0

Note—The variables are, from left to right by column, respondent's identification number, nine indicator variables each for the nine test items, and response categories (1 = correct, 0 = incorrect).

LISTING 2
SAS Statements for Fitting the Rasch Model to the Starter Example

```

DATA starter0;
  INFILE 'starter0.dat';
  INPUT i1-i9;
  person = _N_;
  PROC TRANSPOSE DATA=starter0 OUT=longForm NAME=i PREFIX=score;
  BY person;
  DATA starter;
  SET longForm;
  ARRAY items{9} i1-i9;
  DO j=1 TO 9;
  items{j} = 0;
  END;
  items{SUBSTR(i, 2, 1)} = 1;
  resp = score1;
  DROP i j score1;
  RUN;
  PROC NLMIXED DATA=starter QPOINTS=25;
  PARMS b1-b9=0 sigma=1;
  b = b1*i1+b2*i2+b3*i3+b4*i4+b5*i5+b6*i6+b7*i7+b8*i8+b9*i9;
  eta = theta - b;
  p = 1/(1+EXP(-eta));
  MODEL resp ~ BINARY(p);
  RANDOM theta ~ NORMAL(1.155, sigma*sigma) SUBJECT=person;
  PREDICT p OUT=predProb;
  PREDICT theta OUT=personParm;
  ODS OUTPUT ParameterEstimates=itemParm;
  RUN;

```

Note—The command lines preceding PROC NLMIXED convert the 150 (row) $\times$ 9 (column) data matrix in *starter0* to a data set in *starter* in a long format, as shown in Listing 1.

(Continued on next page)

LISTING 3
SAS Statements for Implementing the Three-Parameter Logistic Model

```

PROC NLMIXED DATA=army QPOINTS=25;
  BOUNDS a1-a4 > 0, c1-c4 > 0, c1-c4 < 1;
  PARMS a1=1.5 a2=.88 a3=1.2 a4=3.3
        b1=-.02 b2=.27 b3=.63 b4=-.74 c1-c4=0.2;
  slope = a1*i1+a2*i2+a3*i3+a4*i4;
  eta = 1.7*slope*(theta - (b1*i1+b2*i2+b3*i3+b4*i4));
  guess = c1*i1+c2*i2+c3*i3+c4*i4;
  p = guess + ((1-guess)/(1+EXP(-eta)));
  MODEL resp ~ BINARY(p);
  RANDOM theta ~ NORMAL(0, 1) SUBJECT=person;
RUN;

```

Note—The data *army* contains dichotomous responses of a sample of 776 respondents arranged in the long format.

LISTING 4
SAS Statements for Implementing the Graded Response Model

```

PROC NLMIXED DATA=capital5 QPOINTS=25;
  BOUNDS d11-d14 > 0, d21-d24 > 0, d31-d34 > 0,
        d41-d44 > 0, d51-d54 > 0 a1-a5 > 0;
  PARMS a1-a5=1 b1-b5=1;
  slope=a1*i1+a2*i2+a3*i3+a4*i4+a5*i5;
  b=b1*i1+b2*i2+b3*i3+b4*i4+b5*i5;
  d1=d11*i1+d21*i2+d31*i3+d41*i4+d51*i5;
  d2=d12*i1+d22*i2+d32*i3+d42*i4+d52*i5;
  d3=d13*i1+d23*i2+d33*i3+d43*i4+d53*i5;
  d4=d14*i1+d24*i2+d34*i3+d44*i4+d54*i5;
  eta1=slope*(theta-b);
  eta2=slope*(theta-(b+d1));
  eta3=slope*(theta-(b+d1+d2));
  eta4=slope*(theta-(b+d1+d2+d3));
  eta5=slope*(theta-(b+d1+d2+d3+d4));
  IF(resp=1) THEN p=1/(1+EXP(-eta1));
  ELSE IF(resp=2) THEN p=1/(1+EXP(-eta1))-1/(1+EXP(-eta2));
  ELSE IF(resp=3) THEN p=1/(1+EXP(-eta2))-1/(1+EXP(-eta3));
  ELSE IF(resp=4) THEN p=1/(1+EXP(-eta3))-1/(1+EXP(-eta4));
  ELSE IF(resp=5) THEN p=1/(1+EXP(-eta4))-1/(1+EXP(-eta5));
  ELSE IF(resp=6) THEN p=1/(1+EXP(-eta5));
  IF(p > 1E-8) THEN ll = LOG(p);
  ELSE ll = -1E100;
  MODEL resp ~ GENERAL(ll);
  RANDOM theta ~ NORMAL(0, 1) SUBJECT=person;
RUN;

```

Note—The data *capital5* contains categorical responses of a sample of 245 respondents arranged in the long format.

LISTING 5
SAS Statements for Implementing the Rating Scale Model for Fitting the Responses
to the Five Statements in the Capital Punishment Data

```

PROC NLMIXED DATA=capital5 QPOINTS=25;
PARMS b1-b5=0 d1-d4=0 sigma=1;
eta=theta-(b1*i1+b2*i2+b3*i3+b4*i4+b5*i5);
num2=EXP(eta-d1);
num3=EXP(2*eta-d1-d2);
num4=EXP(3*eta-d1-d2-d3);
num5=EXP(4*eta-d1-d2-d3-d4);
num6=EXP(5*eta);
denom=1+num1+num2+num3+num4+num5;
IF(resp=1) THEN p=1/denom;
ELSE IF(resp=2) THEN p=num2/denom;
ELSE IF(resp=3) THEN p=num3/denom;
ELSE IF(resp=4) THEN p=num4/denom;
ELSE IF(resp=5) THEN p=num5/denom;
ELSE IF(resp=6) THEN p=num6/denom;
IF(p > 1E-8) THEN ll = LOG(p);
ELSE ll = -1E100;
MODEL resp ~ GENERAL(ll);
RANDOM theta ~ NORMAL(0, sigma**2) SUBJECT=person;
ESTIMATE 'd5' -(d1+d2+d3+d4);
RUN;

```

LISTING 6
SAS Statements for Implementing the Partial Credit Model for Fitting the Responses
to the Five Statements in the Capital Punishment Data

```

PROC NLMIXED DATA=capital5 QPOINTS=25;
PARMS b1-b5=0 d11-d14=0 d21-d24=0 d31-d34=0 d41-d44=0 d51-d54=0
      sigma=1;
beta=b1*i1+b2*i2+b3*i3+b4*i4+b5*i5;
d1=d11*i1+d21*i2+d31*i3+d41*i4+d51*i5;
d2=d12*i1+d22*i2+d32*i3+d42*i4+d52*i5;
d3=d13*i1+d23*i2+d33*i3+d43*i4+d53*i5;
d4=d14*i1+d24*i2+d34*i3+d44*i4+d54*i5;
eta=theta-beta;
num1=EXP(eta-d1);
num2=EXP(2*eta-d1-d2);
num3=EXP(3*eta-d1-d2-d3);
num4=EXP(4*eta-d1-d2-d3-d4);
num5=EXP(5*eta);
denom=1+num1+num2+num3+num4+num5;
IF(resp=1) THEN p=1/denom;
ELSE IF(resp=2) THEN p=num1/denom;
ELSE IF(resp=3) THEN p=num2/denom;
ELSE IF(resp=4) THEN p=num3/denom;
ELSE IF(resp=5) THEN p=num4/denom;
ELSE IF(resp=6) THEN p=num5/denom;
IF(p > 1E-8) THEN ll = LOG(p);
ELSE ll = -1E100;
MODEL resp ~ GENERAL(ll);
RANDOM theta ~ NORMAL(0, sigma**2) SUBJECT=person;
ESTIMATE 'd15' -(d11+d12+d13+d14);
ESTIMATE 'd25' -(d21+d22+d23+d24);
ESTIMATE 'd35' -(d31+d32+d33+d34);
ESTIMATE 'd45' -(d41+d42+d43+d44);
ESTIMATE 'd55' -(d51+d52+d53+d54);
RUN;

```

(Continued on next page)

LISTING 7
SAS Statements for Implementing the Generalized Partial Credit Model for
Fitting the Responses to the Four Items in the Science Data

```

PROC NLMIXED DATA=science QPOINTS=25;
  BOUNDS a1-a4 > 0;
  PARMS a1-a4=1 b1-b4=0 d11=0 d21=0 d31-d32=0 d41-d42=0;
  beta = b1*i1+b2*i2+b3*i3+b4*i4;
  slope = a1*i1+a2*i2+a3*i3+a4*i4;
  d1 = d11*i1+d21*i2+d31*i3+d41*i4;
  d2 = d32*i3+d42*i4;
  eta = theta - beta;
  num1=EXP(slope*(eta+d1));
  num2=EXP(slope*(2*eta)*(i1+i2)+slope*(2*eta+d1+d2)*(i3+i4));
  num3=EXP(slope*(3*eta)*(i3+i4));
  denom1=1+num1+num2;
  denom2=1+num1+num2+num3;
  IF(resp=0 & i1+i2=1) THEN p=1/denom1;
  ELSE IF(resp=1 & i1+i2=1) THEN p=num1/denom1;
  ELSE IF(resp=2 & i1+i2=1) THEN p=num2/denom1;
  ELSE IF(resp=0 & i3+i4=1) THEN p=1/denom2;
  ELSE IF(resp=1 & i3+i4=1) THEN p=num1/denom2;
  ELSE IF(resp=2 & i3+i4=1) THEN p=num2/denom2;
  ELSE IF(resp=3 & i3+i4=1) THEN p=num3/denom2;
  IF(p > 1E-8) THEN ll=LOG(p);
  ELSE ll = -1E100;
  MODEL resp ~ GENERAL(ll);
  RANDOM theta ~ NORMAL(0, 1) SUBJECT=person;
RUN;

```

LISTING 8
Lines of Data Representing the Responses and Codings for Items of the
First and Last Respondents in the Life Satisfaction Survey

1	F	1	1	0	0
1	H	1	0	1	0
1	R	1	0	0	1
1472	F	3	1	0	0
1472	H	3	0	1	0
1472	R	3	0	0	1

Note—The variables are, from left to right by column, respondent's identification number, item (F = family, H = hobbies, R = residence), response categories (1 = *some, a little, or no satisfaction*, 2 = *a fair amount or quite a bit of satisfaction*, 3 = *a great deal or a very great deal of satisfaction*), and three indicator variables for family, hobbies, and residence items, respectively.

LISTING 9
SAS Statements for Implementing the Nominal Categories Model for
Fitting the Responses to a Question on Life Satisfaction

```

PROC NL MIXED DATA=lifesat QPOINTS=30;
PARMS c11=0 c21=0 c31=0 c12=0 c22=0 c32=0 a11=1 a21=1 a22=1;
a12 = a11;
a31 = a21;
a32 = a22;
a1 = a11*i1+a21*i2+a31*i3;
a2 = a12*i1+a22*i2+a32*i3;
c1 = c11*i1+c21*i2+c31*i3;
c2 = c12*i1+c22*i2+c32*i3;
num1 = EXP(a1*theta+c1);
num2 = EXP(a2*theta+c2);
num3 = EXP(-(a1+a2)*theta-c1-c2);
denom = num1+num2+num3;
IF(resp=1) THEN p = num1/denom;
ELSE IF(resp=2) THEN p = num2/denom;
ELSE IF(resp=3) THEN p = num3/denom;
IF(p > 1e-8) THEN ll = LOG(p);
ELSE ll=-1e100;
MODEL resp ~ GENERAL(ll);
RANDOM theta ~ NORMAL(0, 1) SUBJECT=person;
ESTIMATE 'Family c1' c11;
ESTIMATE 'Family c2' c12;
ESTIMATE 'Family c3' -(c11+c12);
ESTIMATE 'Hobbies c1' c21;
ESTIMATE 'Hobbies c2' c22;
ESTIMATE 'Hobbies c3' -(c21+c22);
ESTIMATE 'Residence c1' c31;
ESTIMATE 'Residence c2' c32;
ESTIMATE 'Residence c3' -(c31+c32);
ESTIMATE 'Family a1' -a11;
ESTIMATE 'Family a2' -a12;
ESTIMATE 'Family a3' (a11+a12);
ESTIMATE 'Hobbies a1' -a21;
ESTIMATE 'Hobbies a2' -a22;
ESTIMATE 'Hobbies a3' (a21+a22);
ESTIMATE 'Residence a1' -a31;
ESTIMATE 'Residence a2' -a32;
ESTIMATE 'Residence a3' (a31+a32);
RUN;

```

(Continued on next page)

LISTING 10
SAS PROC IML (Interactive Matrix Language) Statements for Calculating Infit and Outfit Mean-Square and Standardized Statistics for the Rasch Model Fit to the Starter Example

```

PROC IML;

*Read in observed and predicted responses;
*person and item parameter estimates;
USE starter0;
READ ALL VAR{i1 i2 i3 i4 i5 i6 i7 i8 i9} INTO obsResp;
USE predProb;
READ ALL VAR{Pred} INTO predP;
USE personParm;
READ POINT(DO(1,(150-1)*9+1,9)) VAR{Pred StdErrPred} INTO theta;
USE itemParm;
READ POINT{1 2 3 4 5 6 7 8 9} VAR{Estimate StandardError} INTO beta;

nPerson=NROW(obsResp);
nItem=NCOL(obsResp);
nStat=6;

*Compute residuals, variance of residuals, etc;
predResp=SHAPE(predP,nPerson,nItem);
w=predResp#(1-predResp);
w3=predResp##3+(1-predResp)##3;
s=w3/w;
q=w#(w3-w);
sumpw=w[+,];
sumiw=w[+,];
resid=(obsResp-predResp);
sqresid=resid##2;
zSq=sqresid/w;

*Compute item fit statistics;
itemFit=J(nItem,nStat,0);
itemFit[,1]=beta[,1];
itemFit[,2]=beta[,2];
itemFit[,3]=(zSq[+,]/nPerson)`;
varUj=(SQRT(s[+,]/(nPerson##2)-1/nPerson))`;
itemFit[,4]=(itemFit[,3]##(1/3)-1)#(3/varUj)+(varUj/3);
itemFit[,5]=(sqresid[+,]/sumpw)`;
varVj=(SQRT(q[+,]/sumpw))`;
itemFit[,6]=(itemFit[,5]##(1/3)-1)#(3/varVj)+(varVj/3);

*Compute person fit statistics;
personFit=J(nPerson,nStat,0);
personFit[,1]=theta[,1];
personFit[,2]=theta[,2];
personFit[,3]=zSq[+,]/nItem;
varUi=SQRT(s[+,]/(nItem##2)-1/nItem);
personFit[,4]=(personFit[,3]##(1/3)-1)#(3/varUi)+(varUi/3);
personFit[,5]=sqresid[+,]/sumiw;
varVi=SQRT(q[+,]/sumiw);
personFit[,6]=(personFit[,5]##(1/3)-1)#(3/varVi)+(varVi/3);

*Format output;
varName={'Measure','SE','Outfit MNSQ','Outfit ZSTD',
         'Infit MNSQ','Infit ZSTD'};
MATTRIB itemFit ROWNAME=('Item1':'Item9') COLNAME=(varName)
        LABEL={'Item Fit Statistics'} FORMAT=6.2;
PRINT itemFit;
MATTRIB personFit ROWNAME=('Person1':'Person150') COLNAME=(varName)
        LABEL={'Person Fit Statistics'} FORMAT=6.2;
PRINT personFit;
QUIT;

```

Note—Append the code segment to Listing 3.